

NeurIPS 24 Main Track

# Retrieval-Retro: Retrieval-based Inorganic Retrosynthesis with Expert Knowledge

Heewoong Noh, Namkyeong Lee, Gyung S. Na\*, Chanyoung Park\*

Korean Advanced Institute of Science and Technology (KAIST)  
Korea Research Institute of Chemical Technology (KRICT)

\*Corresponding Author

# Background

## Fundamental Goal of Material Science & Material Synthesis

Fundamental Goal of Material Science: Discovering new materials (e.g., semiconductor and batteries)

*How can we establish synthetic routes for newly discovered materials to enable their successful commercialization beyond mere discovery?*

# Background

## Fundamental Goal of Material Science & Material Synthesis

Fundamental Goal of Material Science: Discovering new materials (e.g., semiconductor and batteries)

*How can we establish synthetic routes for newly discovered materials to enable their successful commercialization beyond mere discovery?*



Retrosynthesis Planning

# Background

## Fundamental Goal of Material Science & Material Synthesis

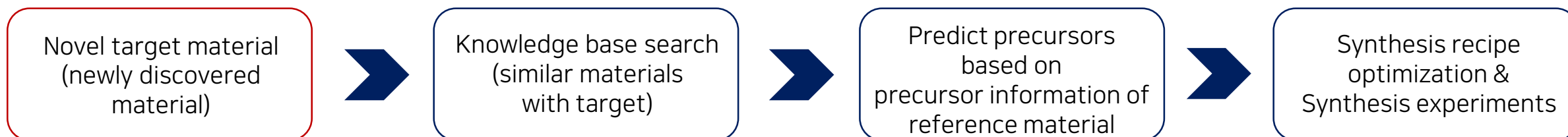
Fundamental Goal of Material Science: Discovering new materials (e.g., semiconductor and batteries)

*How can we establish synthetic routes for newly discovered materials to enable their successful commercialization beyond mere discovery?*



## Retrosynthesis Planning

Conventional approach

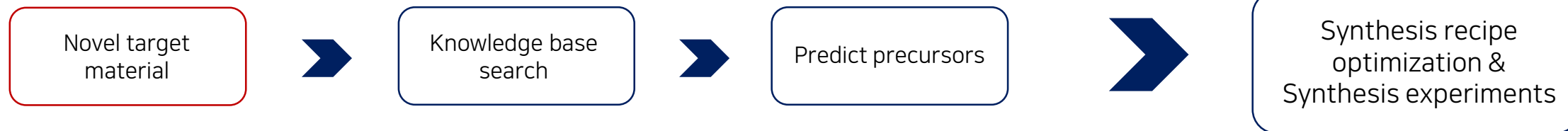


- Identifying similar materials with target materials in the knowledge base
- Rely on chemists' experience and intuition

# Background

## Precursor Prediction

### Conventional approach



# Background

## Precursor Prediction

### Conventional approach

Novel target material



Knowledge base search



Predict precursors



Synthesis recipe optimization & Synthesis experiments

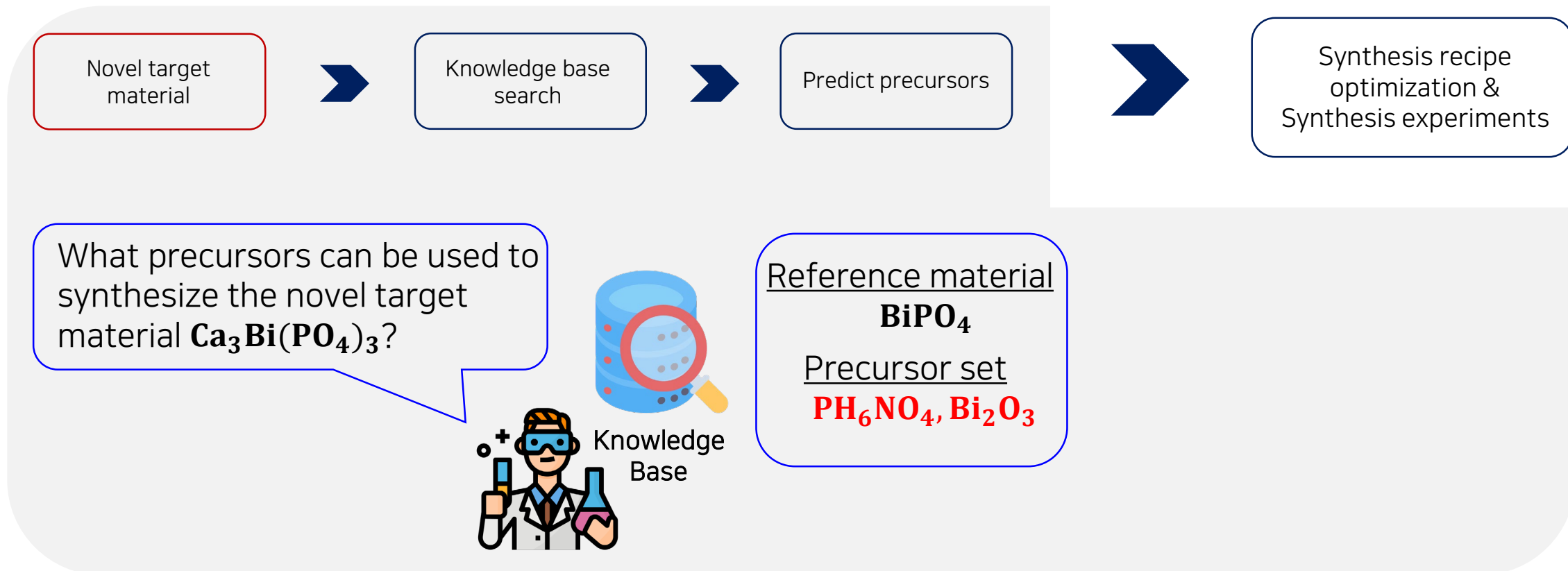
What precursors can be used to synthesize the novel target material  $\text{Ca}_3\text{Bi}(\text{PO}_4)_3$ ?



# Background

## Precursor Prediction

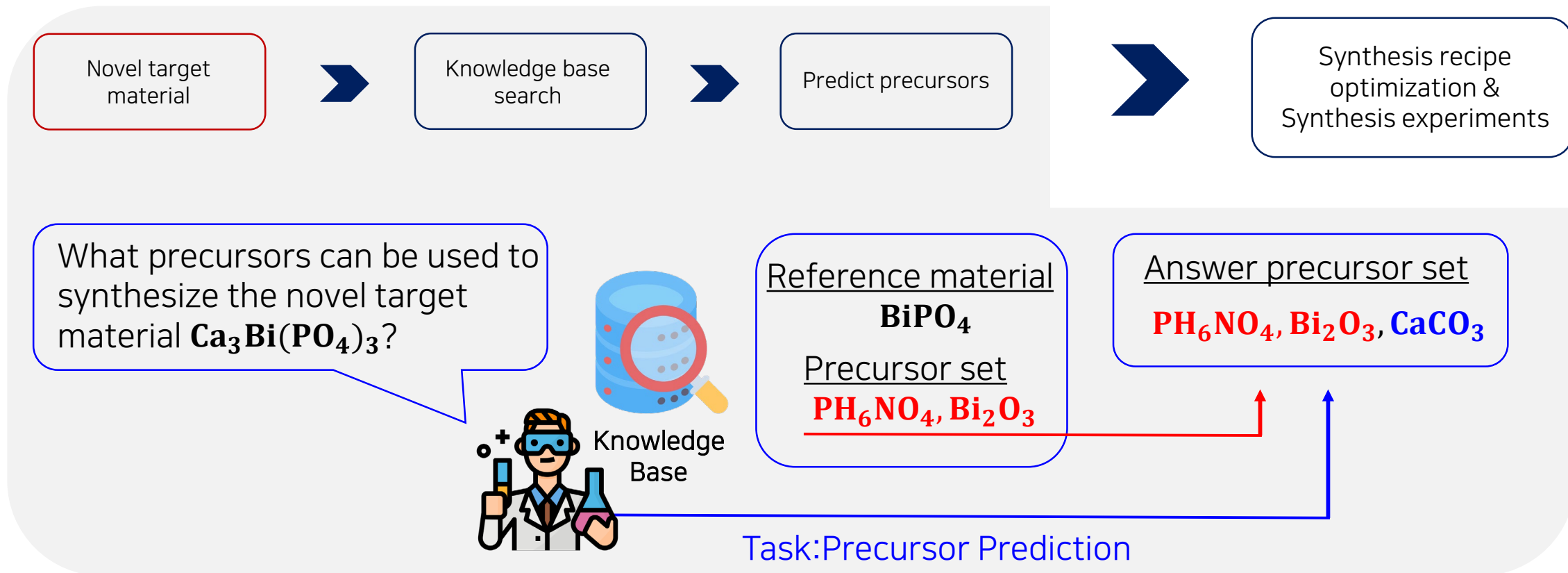
### Conventional approach



# Background

## Precursor Prediction

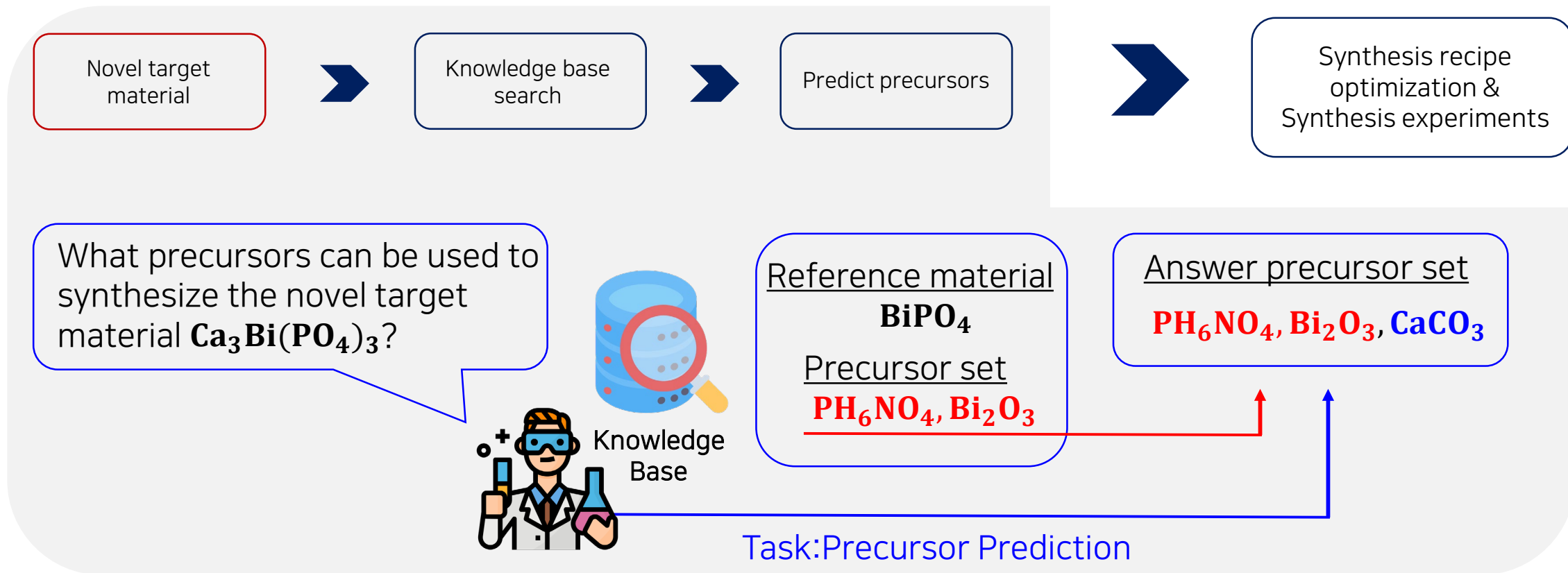
### Conventional approach





# Background Precursor Prediction

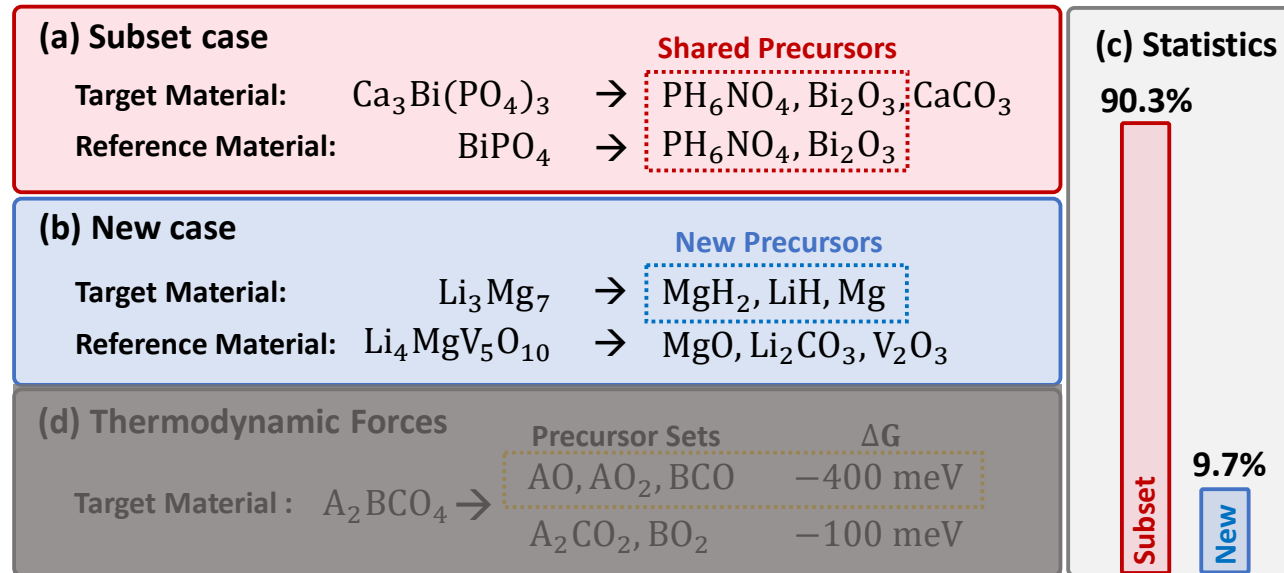
## Conventional approach



To handle novel materials without structural information, we rely solely on chemical composition (i.e., chemical formula)

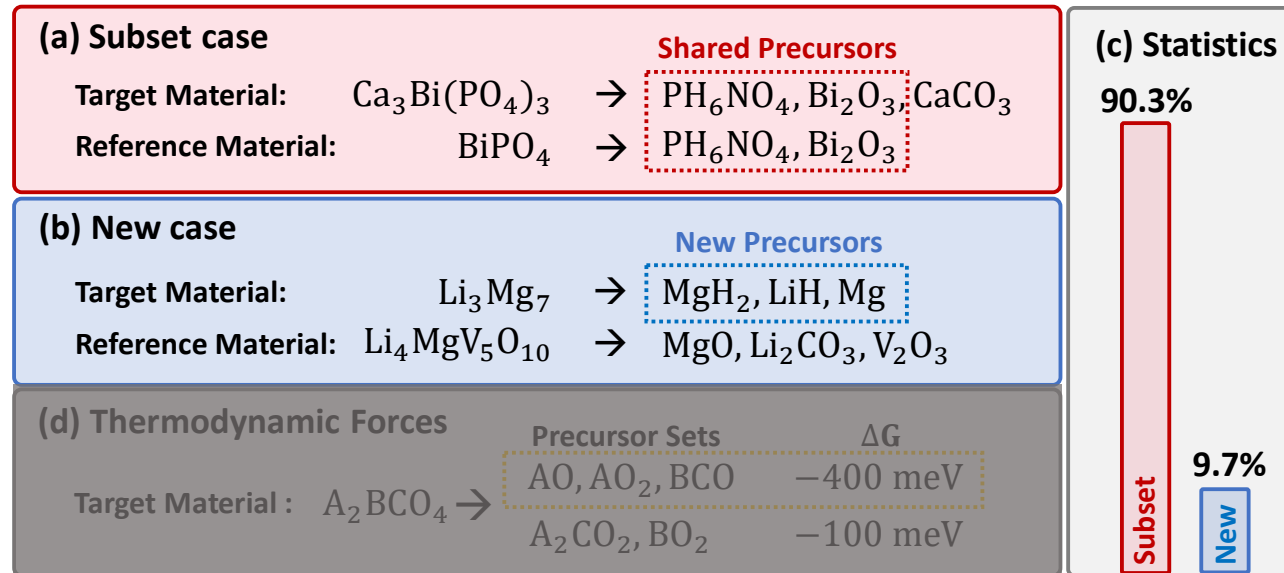
# Motivation

## Discovering Novel Synthesis Recipes



- **Subset case:** Majority of the discovered synthetic routes for the target material share a common set of precursor with the reference material from knowledge base
- **New case:** Entirely new synthesis recipes with new precursor sets → **Novel synthesis recipes**

# Motivation Discovering Novel Synthesis Recipes

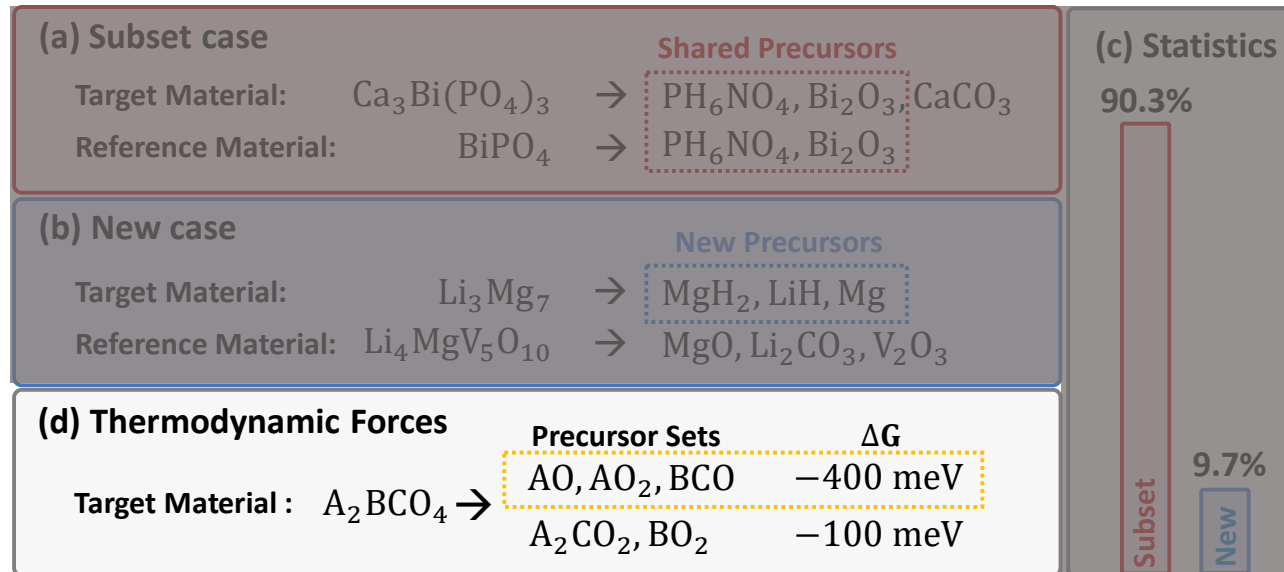


- **Subset case:** Majority of the discovered synthetic routes for the target material share a common set of precursor with the reference material from knowledge base
- **New case:** Entirely new synthesis recipes with new precursor sets → Novel synthesis recipes

Discovering novel synthesis recipes can accelerate the inorganic material synthesis process

# Motivation

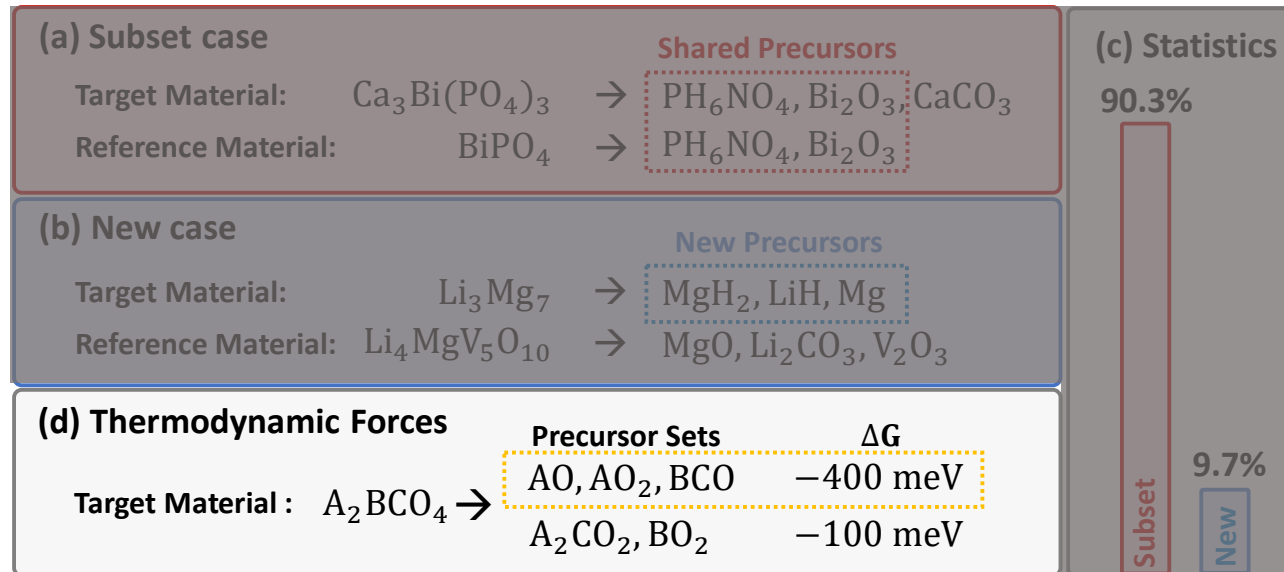
Domain Expertise: Thermodynamic Relationship



- Domain Expertise:** the greater (more negative) thermodynamic driving force ( $\Delta G$ ) between the target material and the precursor set  
 → the more feasible it is to form the target material using the precursor set

# Motivation

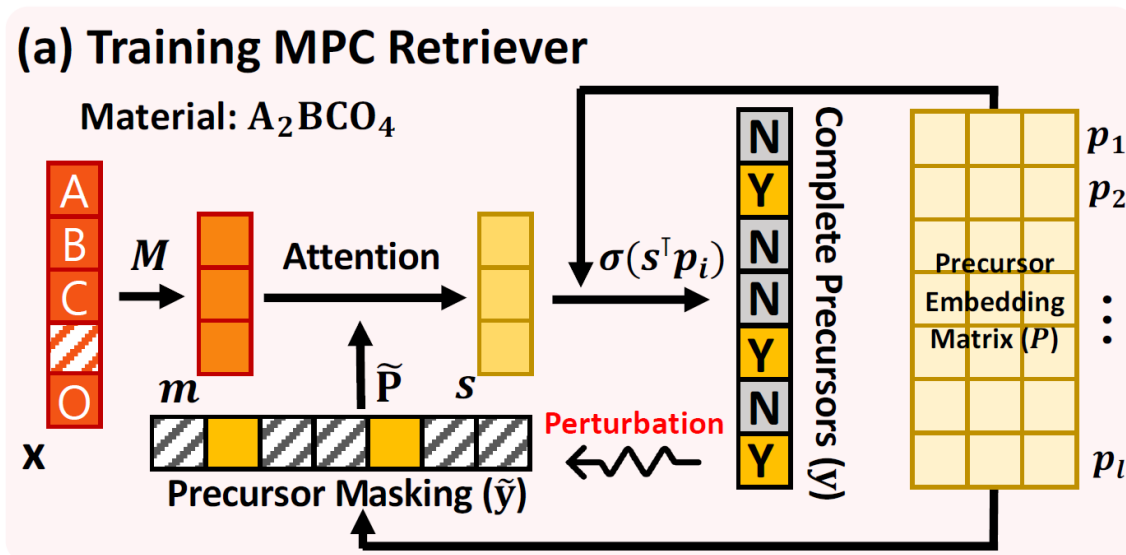
Domain Expertise: Thermodynamic Relationship



- Domain Expertise:** the greater (more negative) thermodynamic driving force ( $\Delta G$ ) between the target material and the precursor set  
 → the more feasible it is to form the target material using the precursor set

Identify effective precursor sets considering thermodynamic driving force  $\Delta G$

# Methodology Reference Material Retrieval: Masked Precursor Completion (MPC) Retriever



Target material embedding:  $\mathbf{m} = M(\mathbf{x})$

Learnable precursor embedding:  $\mathbf{P}$

Perturbed precursor embedding:  $\tilde{\mathbf{P}}$

Probability for each precursor:  $\sigma(\mathbf{s}^T \mathbf{p}_i)$

Trained to reconstruct the original precursor vector

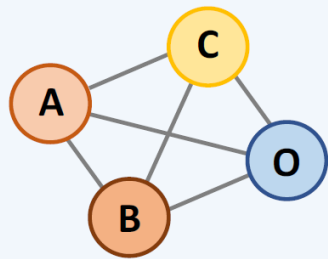
- Following a previous work<sup>1</sup>, we train Masked Precursor (MPC) Retriever
  - Identify reference materials sharing similar precursors with the target material
  - Learn dependencies among precursors and correlation between the precursors and the target material
- Retrieve top-k materials similar to the target material (using  $M$  and cosine similarity)

# Methodology

## Reference Material Retrieval: Neural Reaction Energy Retriever

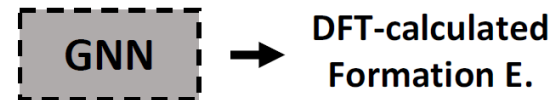
### (b) Training NRE Retriever

Material:  $A_2BCO_4$



Composition Graph  $G$

#### Step1. Pre-train



#### Step2. Fine-tune



$$\Delta G \approx \Delta H = H_{Target} - H_{Precursor\ set}$$

Utilize formation energy of materials (for  $\Delta H$ )  
(Target material and precursor sets)

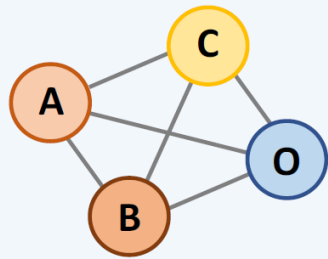
- Thermodynamic driving force between the target material and precursor set can be quantified by  
→ Gibbs free energy ( $\Delta G$ )
- Retrieve materials that have the precursor set capable of inducing favorable reactions with the target material
- $\Delta G$  can be approximated by the difference  $\Delta H$  between the enthalpy of the target and the precursor set

# Methodology

## Reference Material Retrieval: Neural Reaction Energy Retriever

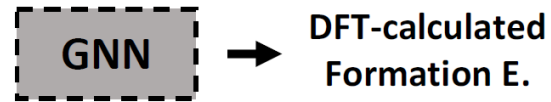
### (b) Training NRE Retriever

Material:  $A_2BCO_4$



Composition Graph  $\mathcal{G}$

Step1. Pre-train



Step2. Fine-tune



Essential to develop formation energy predictor

1) Calculate for any possible material

Composition-based predictor

→ Only use composition without structure

2) Specially designed for experimental data

Pretrain on DFT-calculated formation energy

→ Fine-tune on experimental formation energy

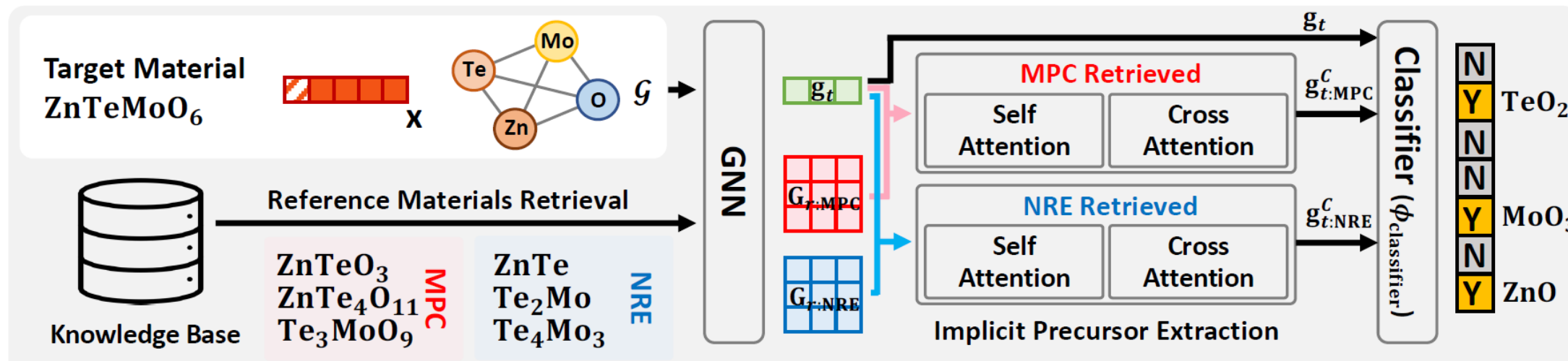
$$\Delta G \approx \Delta H = H_{Target} - H_{Precursor\ set}$$

- Calculate the  $\Delta G$  between target and precursor set, then retrieve  $K$  reference materials that exhibit the most negative  $\Delta G$



# Methodology Implicit Precursor Extraction

**Implicit Precursor Extraction:** Our model does not directly utilize the precursor information of reference materials (i.e., explicit usage); instead, it relies solely on the reference materials themselves



Target Material:  $\mathbf{g}_t$

Reference Material( $K$ ):  $\mathbf{G}_r = [\mathbf{g}_r^1, \dots, \mathbf{g}_r^K]$

Concat:  $\mathbf{G}'_r = [\mathbf{g}'_r^1, \dots, \mathbf{g}'_r^K]$        $\mathbf{g}'_r^k = \phi_1(\mathbf{g}_r^k || \mathbf{g}_t)$

$\mathbf{G}'_r^S = \text{Self-Attention}(\mathbf{Q}_{\mathbf{G}'_r^{S-1}}, \mathbf{K}_{\mathbf{G}'_r^{S-1}}, \mathbf{V}_{\mathbf{G}'_r^{S-1}})$  → [determine which information to extract from the reference materials](#)

$\mathbf{g}_t^C = \text{Cross-Attention}(\mathbf{Q}_{\mathbf{g}_t^{C-1}}, \mathbf{K}_{\mathbf{G}'_r^S}, \mathbf{V}_{\mathbf{G}'_r^S})$  → [learn favorable synthesis recipes from reference materials](#)

$\hat{\mathbf{y}} = \phi_{\text{classifier}}(\mathbf{g}_t || \mathbf{g}_{t:\text{MPC}}^C || \mathbf{g}_{t:\text{NRE}}^C)$

Training Loss:  $\mathcal{L} = -\frac{1}{l} \sum_{i=1}^l [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$

# Methodology

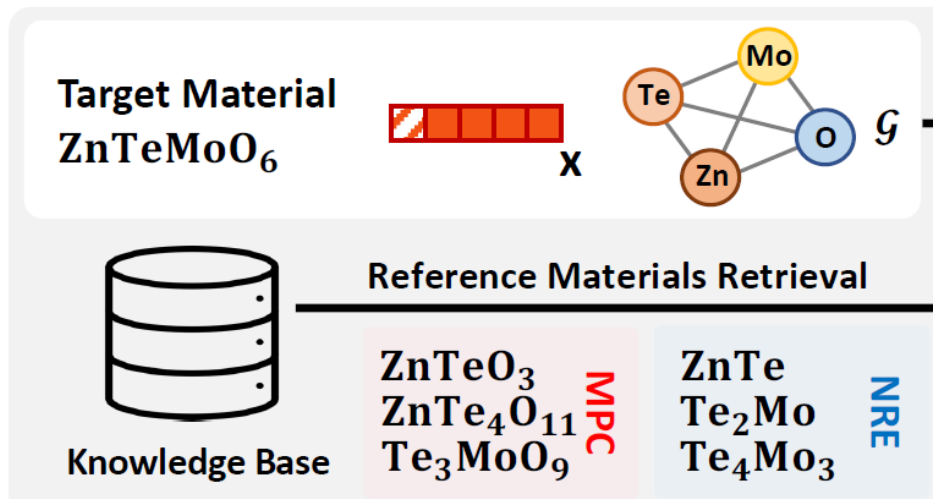
Overall Framework of Retrieval-Retro



1. Target material (composition vector & element fully connected graph)

# Methodology

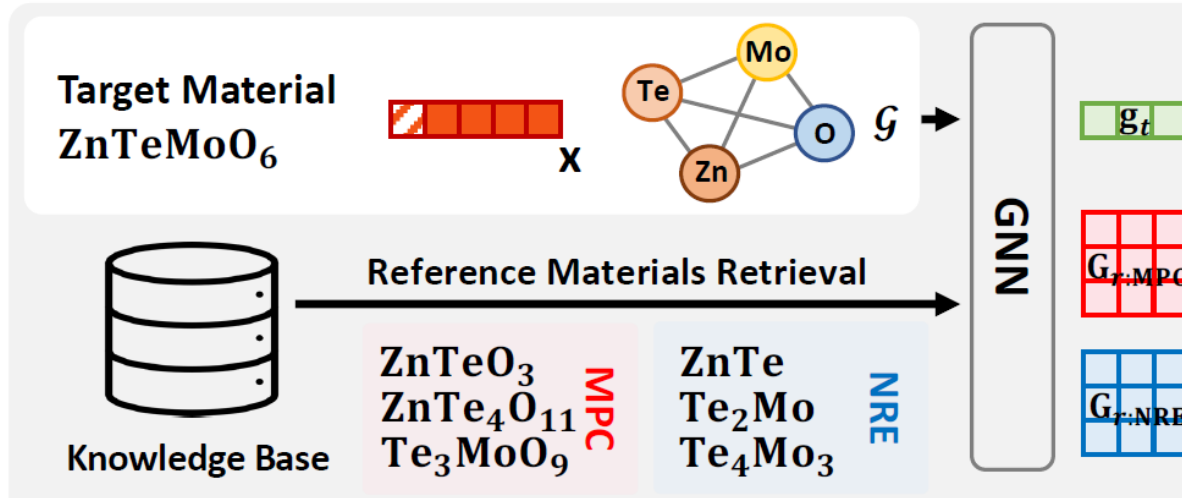
## Overall Framework of Retrieval-Retro



1. Target material (composition vector & element fully connected graph)
2. Retrieve K reference materials from knowledge base using pretrained MPC & NRE retrievers

# Methodology

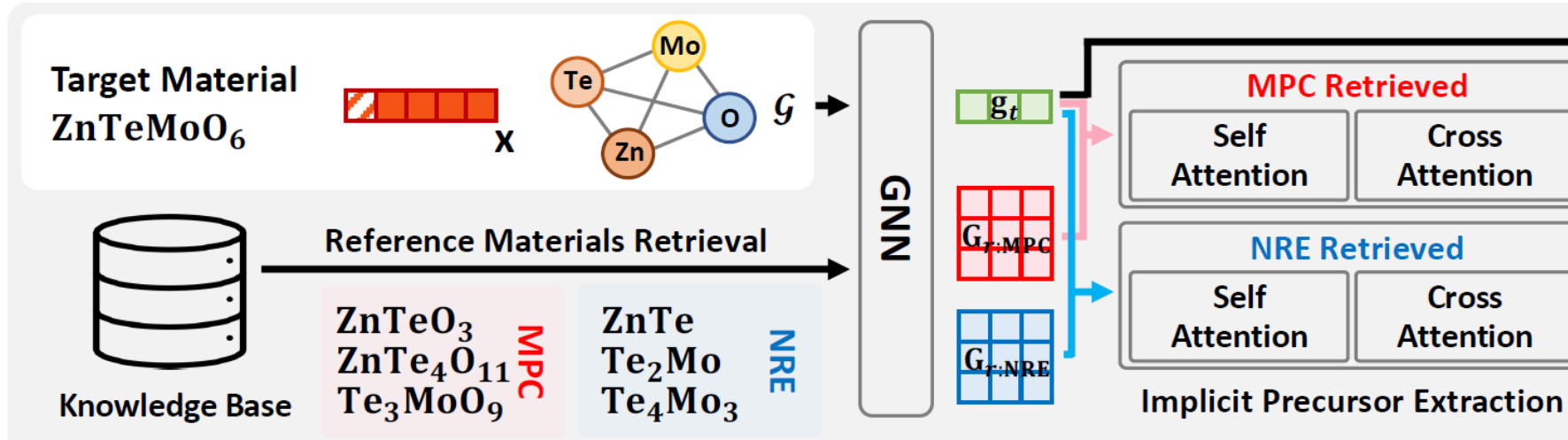
## Overall Framework of Retrieval-Retro



1. Target material (composition vector & element fully connected graph)
2. Retrieve K reference materials from knowledge base using pretrained MPC & NRE retrievers
3. Pass the target material graph and reference material graphs into a GNN

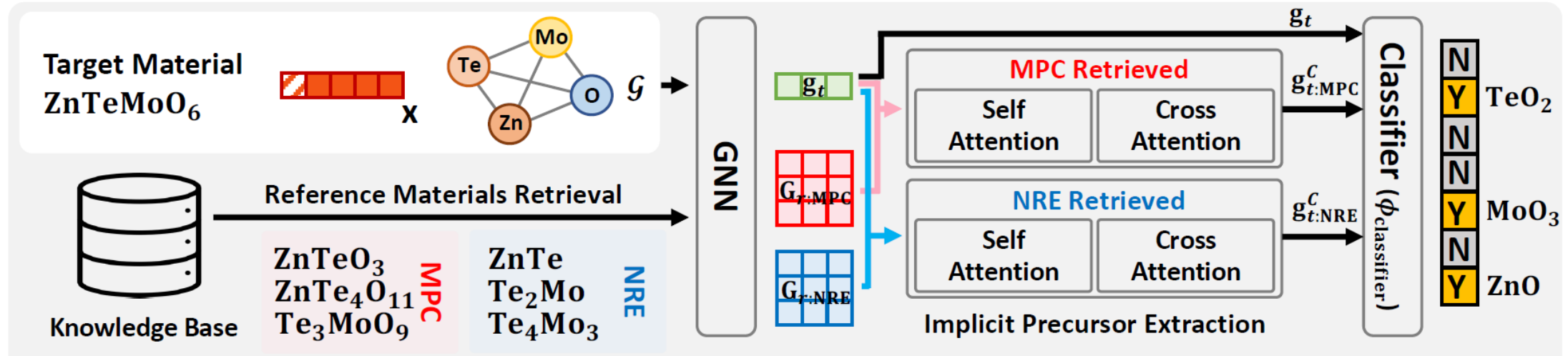
# Methodology

## Overall Framework of Retrieval-Retro



1. Target material (composition vector & element fully connected graph)
2. Retrieve K reference materials from knowledge base using pretrained MPC & NRE retrievers
3. Pass the target material graph and reference material graphs into a GNN
4. Apply self-attention & cross-attention (MPC & NRE branch)

# Methodology Overall Framework of Retrieval-Retro



1. Target material (composition vector & element fully connected graph)
2. Retrieve K reference materials from knowledge base using pretrained MPC & NRE retrievers
3. Pass the target material graph and reference material graphs into a GNN
4. Apply self-attention & cross-attention (MPC & NRE branch)
5. Classifier and calculate the probability of each precursor

# Experiments Dataset & Evaluation Protocol

## Datasets & Evaluation Protocol

33,343 inorganic material synthesis recipes extracted from 24,304 material science papers

Following the preprocessing step, 28,434 target materials are used

**Year-Split:** Train ( ~ 2014) / Valid (2015, 2016) / Test (2017 ~ 2020)

**Random-Split:** Train (80%) / Valid (10%) / Test (10%)

**Knolwedge Base:** Training set

## Importance of Year-Split Setting

Closely replicates the real-world materials discovery conditions

→ Evaluation of the model performance without the need for the costly wet-lab experiments

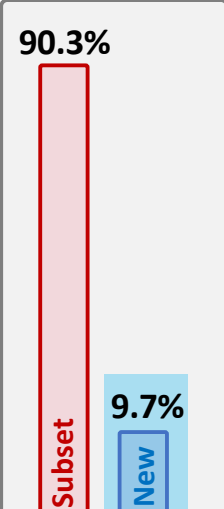
# Experiments Effectiveness of Retrieval-Retro in Inorganic Retrosynthesis Planning

Model	Int.	Retr.	(a) Year Split						(b) Random Split					
			Top-K Accuracy				Recall		Top-K Accuracy				Recall	
			Top-1	Top-3	Top-5	Top-10	Macro	Micro	Top-1	Top-3	Top-5	Top-10	Macro	Micro
Composition MLP	✗	✗	31.60 (1.70)	34.37 (1.58)	35.22 (1.43)	36.56 (0.160)	31.42 (0.030)	31.44 (0.060)	58.56 (0.47)	62.20 (0.36)	62.95 (0.029)	64.32 (0.42)	54.56 (0.44)	55.35 (0.57)
He et al. [12]	✗	✓	45.03 (1.85)	48.02 (1.86)	49.11 (1.77)	51.09 (1.93)	44.72 (1.83)	44.75 (1.86)	61.94 (1.5)	66.44 (1.48)	67.46 (1.55)	68.84 (1.65)	58.55 (1.45)	59.35 (1.34)
ElemwiseRetro	✓	✗	53.45 (0.58)	57.07 (0.52)	58.19 (0.72)	60.84 (0.78)	53.12 (0.60)	53.19 (0.60)	77.23 (0.70)	80.93 (0.54)	81.57 (0.67)	82.78 (0.64)	72.33 (1.14)	73.26 (0.99)
Roost	✓	✗	54.38 (0.75)	57.82 (0.81)	58.82 (1.00)	60.71 (1.15)	54.01 (0.75)	54.04 (0.74)	78.42 (0.91)	82.32 (0.91)	83.07 (0.83)	84.10 (0.66)	73.38 (1.41)	74.46 (1.22)
CrabNet	✓	✗	57.15 (0.77)	61.60 (0.85)	62.44 (0.82)	64.14 (0.86)	56.79 (0.77)	56.82 (0.77)	78.69 (0.78)	81.62 (0.74)	82.27 (0.67)	83.35 (0.56)	73.27 (1.21)	74.28 (0.99)
Graph Network	✓	✗	58.95 (0.41)	63.10 (0.63)	64.07 (0.68)	66.30 (0.62)	58.54 (0.42)	58.61 (0.41)	77.91 (1.31)	81.55 (0.98)	82.37 (0.92)	83.50 (0.90)	72.96 (1.53)	73.88 (1.29)
Graph Network + MPC	✓	✓	<u>60.01</u> (1.10)	<u>64.15</u> (1.10)	<u>65.15</u> (1.17)	<u>67.19</u> (0.83)	<u>59.61</u> (1.10)	<u>59.66</u> (1.10)	<u>79.09</u> (1.25)	<u>82.95</u> (1.13)	<u>83.82</u> (1.19)	<u>84.97</u> (0.94)	<u>73.86</u> (1.34)	<u>74.81</u> (1.23)
<b>Retrieval-Retro</b>	✓	✓	<b>61.16</b> (0.38)	<b>65.92</b> (0.71)	<b>67.18</b> (0.76)	<b>69.45</b> (1.03)	<b>60.97</b> (0.62)	<b>61.06</b> (0.62)	<b>79.81</b> (0.68)	<b>83.62</b> (0.77)	<b>84.46</b> (0.78)	<b>85.70</b> (0.88)	<b>74.61</b> (0.98)	<b>75.49</b> (0.89)

- Modeling interaction among the constituent elements more effective than simple composition vector
- Using precursor information from reference materials from KB enhances the performance
- Retrieval-Retro** surpasses the all baselines, especially for the year split setting, which is more challenging



# Experiments Discovering Novel Synthesis Recipes



Model	Refer.	Retriever		Subset Case				New Case			
		MPC	NRE	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
Graph Network	Explicit	×	×	63.98 (0.34)	67.95 (0.53)	68.83 (0.64)	70.83 (0.81)	16.37 (1.91)	22.00 (3.47)	23.78 (3.48)	27.93 (1.66)
		✓	×	65.01 (1.10)	69.06 (1.17)	69.98 (1.22)	72.03 (0.92)	17.63 (1.66)	22.45 (2.63)	24.22 (2.65)	26.22 (2.74)
Retrieval-Retro	Implicit	✓	×	65.07 (0.80)	69.44 (1.27)	70.41 (1.24)	72.47 (1.46)	19.70 (1.08)	24.52 (1.42)	26.30 (1.28)	30.15 (2.05)
		✓	✓	<b>66.00</b> (0.32)	<b>70.51</b> (0.61)	<b>71.76</b> (0.61)	<b>73.92</b> (0.90)	<b>20.15</b> (1.29)	<b>27.04</b> (1.93)	<b>28.37</b> (2.05)	<b>31.56</b> (3.44)

- Explicitly incorporating MPC retrievers enhances the model performance in Subset Case, however negatively impacts performance in **Top-10 New Case**
- Implicitly integrates precursor information shows performance improvements in both cases  
→ wider performance gap in the **New Case** (a more realistic and challenging scenario)
- NRE retriever consistently enhances the model performance  
→ acquire additional new precursor information that MPC retriever might overlook

# Experiments

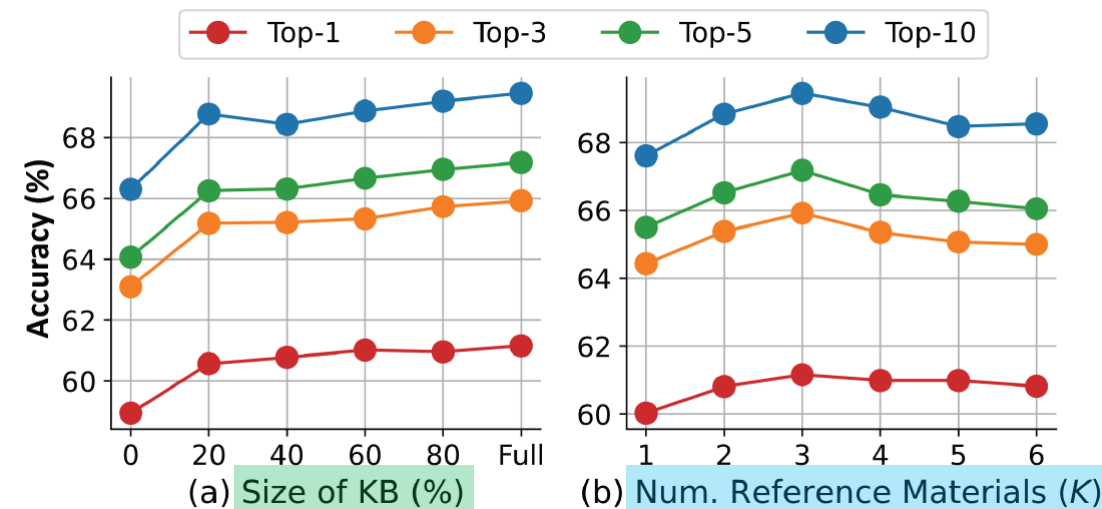
## Model Analysis

Retriever	Top-K Accuracy				Recall	
	Top-1	Top-3	Top-5	Top-10	Macro	Micro
Random	58.42 (1.68)	63.57 (0.64)	64.53 (0.53)	66.61 (0.64)	58.98 (0.64)	59.02 (0.64)
MPC only	59.96 (0.66)	64.49 (0.74)	65.57 (0.96)	<u>68.14</u> (0.81)	59.57 (0.67)	59.64 (0.69)
NRE only	<u>60.28</u> (0.63)	<u>64.70</u> (1.21)	<u>65.75</u> (1.17)	68.00 (1.39)	<u>59.88</u> (0.63)	<u>59.95</u> (0.60)
Retrieval-Retro (MPC + NRE)	<b>61.16</b> (0.38)	<b>65.92</b> (0.71)	<b>67.18</b> (0.76)	<b>69.45</b> (1.03)	<b>60.97</b> (0.62)	<b>61.06</b> (0.62)

- When reference materials are randomly retrieved (Random)
  - Irrelevant precursor information
- Using just one of the retrievers underperforms (i.e., either MPC or NRE)
  - Complementary relationship of MPC and NRE

# Experiments Model Analysis

Retriever	Top-K Accuracy				Recall	
	Top-1	Top-3	Top-5	Top-10	Macro	Micro
Random	58.42 (1.68)	63.57 (0.64)	64.53 (0.53)	66.61 (0.64)	58.98 (0.64)	59.02 (0.64)
MPC only	59.96 (0.66)	64.49 (0.74)	65.57 (0.96)	<u>68.14</u> (0.81)	59.57 (0.67)	59.64 (0.69)
NRE only	<u>60.28</u> (0.63)	<u>64.70</u> (1.21)	<u>65.75</u> (1.17)	68.00 (1.39)	<u>59.88</u> (0.63)	<u>59.95</u> (0.60)
Retrieval-Retro (MPC + NRE)	<b>61.16</b> (0.38)	<b>65.92</b> (0.71)	<b>67.18</b> (0.76)	<b>69.45</b> (1.03)	<b>60.97</b> (0.62)	<b>61.06</b> (0.62)



- When reference materials are randomly retrieved (Random)
  - Irrelevant precursor information
- Using just one of the retrievers underperforms (i.e., either MPC or NRE)
  - Complementary relationship of MPC and NRE

- The larger the KB, the more accurate predictions
- Varying the number of reference materials  $K$ 
  - $K = 3$ , Importance of incorporating precursor information from the reference materials

# Experiments Qualitative Analysis

Target Material:  $\text{Pb}_9[\text{Li}_2(\text{P}_2\text{O}_7)_2(\text{P}_4\text{O}_{13})_2]$

Model	Retriever	Retrieved Material	Corresponding Precursor Sets	Predicted Precursor Set	Answer
Only MPC	MPC	LiNaPbPO Li <sub>0.5</sub> Na <sub>0.5</sub> PO <sub>3</sub> Li <sub>3</sub> V <sub>1.92</sub> Al <sub>0.08</sub> (PO <sub>4</sub> ) <sub>3</sub>	{Li <sub>2</sub> CO <sub>3</sub> , H <sub>3</sub> PO <sub>4</sub> , Na <sub>2</sub> CO <sub>3</sub> , Pb <sub>3</sub> O <sub>4</sub> } {Li <sub>2</sub> CO <sub>3</sub> , NH <sub>4</sub> H <sub>2</sub> PO <sub>4</sub> , NaPO <sub>3</sub> } {Al, V <sub>2</sub> O <sub>5</sub> , LiH <sub>2</sub> PO <sub>4</sub> }	{Li <sub>2</sub> CO <sub>3</sub> , NH <sub>4</sub> H <sub>2</sub> PO <sub>4</sub> }	✗
MPC + NRE (Ours)	MPC	LiNaPbPO Li <sub>0.5</sub> Na <sub>0.5</sub> PO <sub>3</sub> Li <sub>3</sub> V <sub>1.92</sub> Al <sub>0.08</sub> (PO <sub>4</sub> ) <sub>3</sub>	{Li <sub>2</sub> CO <sub>3</sub> , H <sub>3</sub> PO <sub>4</sub> , Na <sub>2</sub> CO <sub>3</sub> , Pb <sub>3</sub> O <sub>4</sub> } {Li <sub>2</sub> CO <sub>3</sub> , NH <sub>4</sub> H <sub>2</sub> PO <sub>4</sub> , NaPO <sub>3</sub> } {Al, V <sub>2</sub> O <sub>5</sub> , LiH <sub>2</sub> PO <sub>4</sub> }	{Li <sub>2</sub> CO <sub>3</sub> , NH <sub>4</sub> H <sub>2</sub> PO <sub>4</sub> , PbO}	✓
	NRE	Pb <sub>3</sub> (PO <sub>4</sub> ) <sub>2</sub> Li <sub>3</sub> P PbP <sub>7</sub>	{PbO, NH <sub>4</sub> H <sub>2</sub> PO <sub>4</sub> } {P, Li} {P, Pb}		

- When only MPC retriever is used, the model fails to predict the entire precursor set
- When the NRE retriever is used with MPC retriever, the model successfully predicts the answer set  
 → can extract precursor information from  $\text{Pb}_3(\text{PO}_4)_2$ , which has the essential precursor **PbO**

# Conclusion

- We propose **Retrieval-Retro**, a novel approach for inorganic retrosynthesis planning that implicitly extracts precursor information from retrieved reference materials.
- **Retrieval-Retro** employs multiple attention layers to enhance and extract relevant information from reference materials.
- **Retrieval-Retro** integrates precursor information from a diverse range of reference materials, supported by the complementary assistance of a neural reaction energy (NRE) retriever designed to leverage expert knowledge.
- Extensive experiments, including realistic scenarios, demonstrate the effectiveness and superiority of **Retrieval-Retro** in discovering novel synthesis pathways for target materials.

# Thank you!

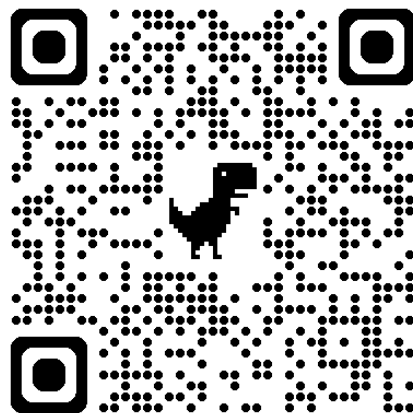
[Full Paper] <https://arxiv.org/abs/2410.21341>

[Source Code] <https://github.com/HeewoongNoh/Retrieval-Retro>

[Lab Homepage] <http://dsail.kaist.ac.kr>

[Email] [heewoongnoh@kaist.ac.kr](mailto:heewoongnoh@kaist.ac.kr)

## Paper



## Code

